

DOCUMENT RESUME

ED 189 152

TM 800 322

AUTHOR Frary, Robert E.; Giles, Mary B.
TITLE Multiple-Choice Test Bias Due to Answering Strategy Variation.
INSTITUTION Virginia Polytechnic Inst. and State Univ., Blacksburg.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE Apr 80
GRANT NIE-G-79-0140
NOTE 36p.: Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Achievement Tests; Analysis of Variance; Goodness of Fit; *Guessing (Tests); Item Analysis; Junior High Schools; Latent Trait Theory; *Mathematical Models; *Multiple Choice Tests; Racial Differences; Sex Differences; *Test Bias; Testing Problems; *Test Wisdom

ABSTRACT

This paper describes the development and investigation of a new approach to determining the existence of bias in multiple-choice test scores. Previous work in this area has concentrated almost exclusively on bias attributable to specific test items or to differences in test score distributions across racial or ethnic groups. In contrast, the present work concentrates ultimately on the behavior of the individual examinee in choosing among the responses to each item. The major approach was to investigate variation in person-fit, that is, the extent to which an individual examinee selects the choices which are more probable based on the examinee's score and the overall popularity of the choices in the population. Four different person-fit measures were computed for random samples representing subpopulations defined by race, sex, score range and geography. There was strong evidence of variation in person-fit between black and white and between male and female examinees, especially in the lower score ranges. However, subsidiary analyses suggested that this outcome did not reflect bias in the scores from the tests analyzed. Nevertheless, the person-fit approach was shown sensitive to group differences and could reveal the presence of bias if it exists. (Author/CTM)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. FRARY

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

MULTIPLE-CHOICE TEST BIAS
DUE TO
ANSWERING STRATEGY VARIATION

by

Robert B. Frary

Abstract

This paper describes the development and investigation of a new approach to determining the existence of bias in multiple-choice test scores. Previous work in this area has concentrated almost exclusively on bias attributable to specific test items or to differences in test score distributions across racial or ethnic groups.

In contrast, the present work concentrates ultimately on the behavior of the individual examinee in choosing among the responses to each item. A number of statistical analyses consistent with this goal were applied to responses to two statewide test batteries. The major approach was to investigate variation in person-fit, that is, the extent to which an individual examinee selects the choices which are more probable based on the examinee's score and the overall popularity of the choices in the population. Four different person-fit measures were computed for random samples representing subpopulations defined by race, sex, score range and geography.

There was strong evidence of variation in person-fit between black and white and between male and female examinees, especially in the lower score ranges. However, subsidiary analyses suggested that this outcome did not reflect bias in the scores from the tests analyzed. Nevertheless, the person-fit approach was shown sensitive to group differences and could reveal the presence of bias if it exists.

MULTIPLE-CHOICE TEST BIAS
DUE TO
ANSWERING STRATEGY VARIATION¹

by

Robert B. Frary

Mary B. Giles

Virginia Polytechnic Institute
and State University
Blacksburg, Virginia

This paper describes the development and investigation of a new approach to determining the existence of bias in multiple-choice test scores. Previous work in this area has concentrated almost exclusively on bias attributable to specific test items or to differences in test score distributions across racial or ethnic groups.

Analytically, a test may be defined as biased if a definable subgroup of examinees has an average score lower than that of another but performs as well (or better) on an independent, direct measure of an appropriate criterion. Many deficiencies have been reported in traditional procedures for investigating bias based on this definition (e.g. Cole, 1973; Linn and Werts, 1971).

The problems of analysis based on external criteria may be avoided by investigating bias at the item level without reference to external criteria. By one definition, a test is considered biased if there is a significant group by item interaction as determined by analysis of variance (Cleary and Hilton, 1968). However, the significance tests from this procedure are probably not valid since the underlying assumptions of analysis of variance are not met (Echternacht, 1974).

Another approach toward detecting bias, either at the item or total score level, without reference to external criteria is provided by item characteristic curve theory. This approach (see Hambleton, et al., 1977) has many advantages, the most important of which is that item difficulty estimates can be made independently of the ability distribution of the sample. This characteristic is essential for an objective approach to detecting bias. For example, it is well known that black examinees typically attain average scores below those of whites. If these differences reflected differences in ability, comparison of percentage-right scores would not

¹The research reported in this paper was supported by the National Institute of Education, Grant No. NIE-G-79-0140, Project No. 9-0498.

determine item bias. However, item characteristic curve theory examines item differences across groups statistically unaccounted for by differences in ability. Items that exhibit large residuals (that is, differences unaccounted for by ability) across groups are considered biased items.

In practice, however, application of item characteristic curve theory has not yielded a definitive resolution to the perceived problem of test bias. Analyses of various tests have identified specific items as biased, but subsequent content analysis has usually failed to yield consistent explanations as to why the bias exists (see Lane, 1978). Test developers are therefore no better off than previously in their efforts to write unbiased items.

In contrast to these approaches which deal primarily with specific test items or differences in test score distributions, the research reported here concentrates ultimately on the behavior of the examinee in choosing among the responses to each test item. To see how differences in response selection behavior could cause bias, consider the following hypothetical example. Two examinees each know with assurance the answers to 10 questions on a 40-item, 5-choice test. Examinee A guesses at random on the remaining 30 items and gains the six additional points expected under average luck. Examinee B eliminates an average of two choices per item before guessing among the remaining three and gains an additional 10 points. The resulting scores of 16 and 20 may be sufficiently different to affect Examinee A adversely. Yet, there is substantial reason to believe Examinee A could have successfully undertaken Examinee B's elimination process, since they knew the same number of answers with assurance. Regardless, had the test been of the short-answer variety, the scores would probably have been equal, and, surely, ability to cope with the multiple-choice response mode could seldom seriously be promoted as an appropriate factor in determination of test scores. Thus, the test in question was biased against Examinee A, not because of its content, but because Examinee A did not, for whatever reason, eliminate choices prior to guessing. This characteristic may have been culturally induced or even have been irrational on Examinee A's part. Nevertheless, the testing outcome was unfair in that it wrongly labeled Examinee A as less competent than Examinee B in the area tested, though, admittedly, not in ability to take multiple-choice tests.

Speaking more generally, response behavior systematically different from that of typical examinees could be identified by summing over items the extent to which an examinee's responses differ from expectation. At a crude level, the person-fit statistic recommended by Wright, et al., (1976) for use with the Rasch model serves this purpose. This statistic is:

$$\sum_{i=1}^N [(p_i - E p_i)^2 / (E p_i (1 - E p_i))]. \quad (1)$$

Summation is over items for an individual examinee, and $p_i = 1$ if the

examinee answers item i correctly and 0 otherwise. Ep_i is the probability of correctness for the examinee as determined by the Rasch model (see Wright, 1977). Formula (1) is χ^2 distributed with $N - 1$ degrees of freedom.

While this statistic has not been widely used or tested, it does offer an alternative approach to investigating bias. Individuals who mark items incorrectly they should get right and vice-versa are the ones who generate high values of (1). This outcome could be due either to some inappropriate characteristic of the test or to some interaction between the test and the individual. Thus, the search of causes of bias shifts from the test items and content to the individual, who may be interviewed or otherwise investigated for possible causal factors.

New Person-Fit Measures

A basic weakness of the Rasch person-fit statistic is that on a multiple-choice test information as to which wrong choice was selected is ignored. Only the fact that an item was marked either correctly or incorrectly is taken into consideration. For example, it might be quite probable for an examinee to miss a certain item, but very improbable for a certain wrong choice to be selected. The Rasch person-fit statistic makes no use of this information.

To alleviate this weakness, statistics may be defined based on a matrix of probabilities unique to each examinee, specifically $\bar{p}_{ij}$, the matrix of probabilities that the (unique) examinee will select choice j of item i . These probabilities may be easily estimated by a method presented in Frary, Tideman and Watts (1977), based on the proportion selecting each response and examinee score. Then a person-fit statistic based on $\bar{p}_{ij}$ is:

$$\sum_{i=1}^N [(p_{i*} - Ep_{i*}) / \sqrt{Ep_{i*}(1 - Ep_{i*})}]^2, \quad (2)$$

in which p_{i*} is the probability associated with the actual choice of the examinee for item i . Ep_{i*} is the expected value of p_{i*} , and is estimated as:

$$Ep_{i*} = \sum_j p_{ij}^2, \quad (3)$$

which is simply the weighted average of the p_{ij} 's for item i . By the central limit theorem, formula (2) should be distributed approximately normally

with a mean of 0 and a standard deviation of 1 for tests of sufficient length.

Because the fit for each item is standardized before summation, formula (2) effectively equalizes the relative contribution of each item to the sum. A statistic without this characteristic is obtained by summing across items the value, $p_{i_{\max}} - p_{i\star}$, in which $p_{i_{\max}}$ is the largest probability for any response to item i by the examinee in question. The lower this sum, the more the examinee conformed to the more probable choices. The standardized statistic would be;

$$\frac{[\sum_i (p_{i_{\max}} - p_{i\star}) - E\sum_i (p_{i_{\max}} - p_{i\star})] / \sqrt{V\sum_i (p_{i_{\max}} - p_{i\star})}} \quad (4)$$

The expected value of the sum is easily estimated as:

$$E\sum_i (p_{i_{\max}} - p_{i\star}) = \sum_i (\sum_j p_{ij} (p_{i_{\max}} - p_{ij})) , \quad (5)$$

which is simply the average of the possible differences with each weighted by the probability of its occurrence.

The variance of the sum, $\sum_i (p_{i_{\max}} - p_{i\star})$, may be obtained by adding the terms of the corresponding variance-covariance matrix. The sum of variances may be estimated as follows:

$$\sum_i s_i^2 = \sum_i (\sum_j p_{ij} [Ed_i - (p_{i_{\max}} - p_{ij})]^2) , \quad (6)$$

in which Ed_i is the expected difference for item i as summed in (5). Thus, (6) is the sum across items of the weighted averages across choices of the squared differences between the observed and expected probability differences. The covariances may be expressed in a similar manner and can easily be shown to sum to zero if it is assumed that the items are independent. Specifically, this assumption requires that, for a specific examinee,

$$p_{ijkh} = p_{ij} p_{kh} , \quad (7)$$

in which p_{ijkh} is the probability of choosing choice j of item i and choice h of item k . While this assumption is not strictly true for any given pair

of choices, it is very close to the truth as evidenced by previous research in this area (see Frary, Tideman and Watts, 1977). Thus it is reasonable to use (6) to estimate the variance in (4).

Professor I. J. Good, of the Department of Statistics at Virginia Polytechnic Institute and State University, has also considered the problem of unusualness of an examinee's responses and recommends use of a "surprise" index proposed by Weaver (1948) and generalized by Good (1956). Specifically, for each examinee, compute:

$$\Lambda = \sum_i [\sum_j p_{ij} \log p_{ij} - \log p_{i*}] . \quad (8)$$

As shown by Good, the expected value of this statistic is 0, and its variance is:

$$V[\Lambda] = \sum_i [\sum_j (\log p_{i*})^2 - (\sum_j p_{ij} \log p_{ij})^2] .$$

The three person-fit statistics just established, (2), (4), and (8), and Rasch person-fit measure previously described (1) are those which were used in the study and are referred to hereafter as follows:

- (1) Method A - the Rasch person-fit measure.
- (2) Method B - standardization before summation over items.
- (3) Method C - standardization after summation over items.
- (4) Method D - logarithmic "surprise" index by Good.

Item Statistics

The statistics just developed for person-fit across all items are such that a small proportion of items strongly biased against certain groups could result in high person-fit statistics for members of those groups. In this case, the efforts to verify bias would substantially overlap conventional methods of detecting bias arising from specific items. On the other hand, existing bias could affect responses to all items, rendering the usual item-bias methods useless. Ability to detect bias under this circumstance is the main strength of the statistics proposed.

In any event, it was judged desirable to check for the conventional form of item bias for the data of this study. Toward this end, the following two statistics were defined for each test item, i:

$$\sum_e (Ep_{i*} / p_{i*}) / N , \quad (9)$$

$$\sum_e (Ep_{i*} - p_{i*}) / N , \quad (10)$$

in which summation is over examinees and N is the number of examinees. These reflect as ratios and differences respectively the extent to which the average examinee's choice differs from expectation. In addition the Rasch fit mean square (wee Wright, 1977) was adopted for evaluating item characteristics.

Another item statistic relevant to the goals of this investigation is the index of dispersion,

$$\left(\sum_{j=1}^M |f_j - \bar{f}_i| \right) / M , \quad (11)$$

in which summation is over wrong choices, where f_j is the proportion of examinees selecting wrong choice j and $\bar{f}_i = (\sum f_j) / M$ (or $\bar{f}_i$ is the average proportion of examinees selecting a wrong choice on item i). This statistic may be averaged across items for comparison of different tests. A low value would indicate a tendency for examinees to guess at random when an answer is unknown.

A method proposed by Horst (1966) was used to estimate the proportion of examinees who answer an item correctly without guessing. Under Horst's model, it is assumed that those selecting the least popular wrong choice do so by guessing at random, that the answer will be the most popular choice and that differences in response proportions across the remaining wrong choices reflect the ability to eliminate increasing numbers of wrong choices prior to guessing among the remaining choices. The result is that the (nonnegative) difference between the proportion selecting the answer and the proportion selecting the most popular wrong choice is the proportion who knew the answer without guessing. If this difference is negative, it must be concluded that no examinee knew the answer and that some excluded the right answer prior to guessing. Of course this model may be much more appropriate for some items than others. Nevertheless, summing over items the estimated proportions of examinees with sure knowledge should yield a measure which would permit valid comparisons between groups if not a very accurate measure of the average score component gained without guessing.

Data Sources

The data were derived from two sources. The first data set consisted of responses from approximately 24,000 eighth graders to a standardized achievement test battery administered as part of a state-wide testing program. The battery was composed of eight tests: Test 1 - Reading Vocabulary; Test 2 - Reading Comprehension; Test 3 - Reading Essential Skills; Test 4 - Study Skills; Test 5 - Math Computation; Test 6 - Math Problem Solving; Test 7 - Math Essential Skills; Test 8 - Occupational Information. This set of tests is referred to subsequently as the eight-test battery (ETB).

The second data set was made up of approximately 86,000 responses from ninth graders to the spring, 1979, administration of the Virginia Graduation Competency Tests (VGC). This battery consisted of a reading test (60 items) and a mathematics test (99 items). The reading section requires interpretation of various types of labels and directions, form completion, use of references, determining main ideas from short passages, and understanding simple documents. The mathematics test assesses facility with basic numerical operations, use of charts and graphs, practical geometry, and consumer problems. These tests are similar respectively, to the Reading Essential Skills and to the Math Computation/Math Essential Skills tests of the ETB.

Procedures and Results

Sampling

Similar computational and analytical procedures were followed for both test batteries. The first step in the analysis was to perform total score and item analyses on each of the individual tests for the two groups of examinees. KR20 reliability coefficients and average test scores for the total group, for white examinees, and for black examinees were determined for each test. These results are shown in Table 1. The mean test scores for white examinees were, for all tests, higher than the mean scores for black examinees and, while the standard deviations were not always the same for whites and blacks, there was no consistent pattern in the ETB of one racial group being more heterogeneous than the other. On the VGC, however, the standard deviations for the blacks were larger than those for the whites, particularly on the mathematics test, and, since the means of the tests were quite high, the larger standard deviations indicate more scores for black examinees in the lower ranges.

In addition to means, standard deviations and reliabilities, the item analysis program provided the proportions of examinees selecting each of the options for each item on each of the tests. These proportions were used in later analyses to represent the probabilities of the selection of the related options for an average examinee and are referred to subsequently as the item probabilities. The point-biserial correlation between each choice and total score was also provided by the item analysis program. These statistics were used to determine whether any items should be deleted from subsequent analyses, since a negative correlation for a correct choice or a substantially positive correlation for a wrong choice would indicate violation of assumptions on which the person-fit statistics are based.

Table 1: DESCRIPTIVE STATISTICS FOR TOTAL GROUP, WHITES, AND BLACKS
EIGHT-TEST BATTERY

	KR20	Total Group			Whites			Blacks		
		$\bar{x}$	s.d.	N	$\bar{x}$	s.d.	N	$\bar{x}$	s.d.	N
Test 1	.92	24.45	9.04	22,904	26.71	8.21	17,475	15.64	6.81	3,723
2	.82	13.10	4.25	24,531	14.18	3.70	18,317	9.32	3.90	4,366
3	.93	36.52	8.04	24,193	38.60	6.39	18,160	29.01	9.24	4,200
4	.81	11.79	4.36	23,437	12.84	4.03	17,589	8.02	3.40	4,066
5	.94	32.22	10.92	21,040	34.33	10.26	16,114	23.28	9.37	3,364
6	.89	16.75	6.89	23,413	18.47	6.37	17,614	10.48	4.99	4,048
7	.87	18.52	5.47	24,077	19.88	4.78	18,081	13.39	4.98	4,194
8	.90	28.51	7.60	23,470	30.51	6.30	17,761	20.84	7.78	3,938

VIRGINIA GRADUATION COMPETENCY TESTS

	KR20	Total Group			Whites			Blacks		
		$\bar{x}$	s.d.	N	$\bar{x}$	s.d.	N	$\bar{x}$	s.d.	N
Reading	.89	48.60	7.94	77,721	50.32	6.91	59,982	43.36	8.72	18,352
Math	.87	85.18	8.68	79,697	87.52	6.99	54,941	77.44	13.99	18,370

Because of the possibilities of exaggerated effects from omissions in the person-fit statistics from Methods B, C and D, examinees omitting more than 10% of the items in a given test in the ETB were excluded. A far larger percentage of black examinees omitted more than 10% of the test items than did white examinees, and a slightly larger percentage of males of both races omitted more than 10% of the items than did females of corresponding races. Depending on the test, between 8% and 30% of the black examinees were excluded from the analyses on the ETB because of omissions and between 4% and 17% of white examinees.

Since the number of examinees omitting each item on the VGC was fewer than 1,000 (about 1%) for the reading test and fewer than 200 (far less than 1%) for the mathematics test, examinees with any omissions were excluded from the population from which the samples were drawn. To do otherwise might have yielded highly inflated person-fit statistics under Methods B, C and D, for examinees with only one or two omissions. Also excluded were handicapped examinees and those who were retaking one or both of the tests. Approximately 14.0% of black examinees were dropped because of omissions and 7.3% of white examinees.

The next step in the analysis was to select random samples of approximately 300-400 from the acceptable cases for each test in each test battery. The SAMPLE procedure from the Statistical Package For Social Scientists (SPSS, 1975) was used for this purpose. Some difficulties were encountered with the SAMPLE procedure, however. In this procedure, it is necessary to request a proportion of the total group rather than the actual number of cases desired. When the exact proportion was entered into the program, the resulting numbers of cases selected varied considerable because of the internal workings of the program over which no control could be exerted. Sometimes too few cases and other times too many cases would be selected. Because of this problem, it was decided to err on the abundant side, and so the N's of the samples tend to be larger than the original 300 cases planned for each random sample.

The means and standard deviations for the random samples are reported in Table 2. The similarities between these statistics for the samples and the total group suggest that the samples are representative of the test characteristics of the total groups. Comparisons of the distributions of race and sex in the total groups and in the random samples were also made. The percentages of black and white and male and female examinees in the random samples were found to be highly similar to the percentages in the total group. The greatest deviation of the sample percentage from the total group was found in Test 7 of the ETB in which the percentages of white and black examinees in the total group were, respectively, 75% and 17%, and, in the random sample, 71% and 21%. On the ETB, approximately 75% of the examinees were white and 16% to 21% were black. Approximately 7% of the examinees in this group were of other racial groups or failed to code race on their answer sheets. On the VGC, less than 2% were of other racial extraction or omitted the race designation. Overall, in the Virginia group approximately 75% of the examinees were white and 23% black. The distribution of sex was approximately 50-50 in both the ETB and VGC, with slightly more males included in most of the tests in the ETB.

Table 2: DESCRIPTIVE STATISTICS FOR RANDOM SAMPLES
EIGHT-TEST BATTERY

Total Test Score			
	$\bar{x}$	s.d.	N
Test 1	24.22	9.80	312
2	12.90	4.20	461
3	35.97	8.32	480
4	11.71	4.16	457
5	31.36	10.34	400
6	16.74	7.13	481
7	18.08	5.23	445
8	28.91	7.66	445

VIRGINIA GRADUATION COMPETENCY TESTS

Total Test Score			
	$\bar{x}$	s.d.	N
Reading	49.67	7.45	365
Mathematics	86.44	10.55	380

Person-Fit Statistics

After the random samples were drawn and defined, the four person-fit statistics were calculated for each of the tests. The means and standard deviations for these person-fit statistics are in Table 3. The expected mean value of the Method A statistic is the number of items in the test from which it is computed, e.g. for Test 1 the expected mean is 40. For the remaining person-fit statistics the expected mean is zero. For all tests and all person-fit statistics in the ETB, the computed means for the random samples are less than would be expected theoretically. However, two of the mean VGC reading test person-fit statistics (Methods C and D) are slightly positive, in contrast to the negative values found for the same statistics on the ETB.

The degree of overlap among the four person-fit statistics and their dependency upon total test scores were examined by computing Pearson correlation coefficients among these statistics and the total test score for each test. These are shown in Table 4. Method A is more closely related to Method B than to Methods C and D. Methods B, C, and D show considerable overlap. For most of the tests, the relationships between the person-fit statistics and total test scores was negative. This is the expected result since high scorers would choose mostly right answers and have little opportunity to select unexpected options, producing low person-fit statistics. The reverse would be true of examinees with low test scores. An exception, however, occurred for Test 5 (Math Computation) and the VGC math test. For these tests, the correlations with total test scores were mostly positive, suggesting that the unusual responses tended to be made by high scorers rather than low. The relationships between the person-fit statistics and total test scores were generally less strong in the VGC than in the ETB.

The next step in the analysis was to select samples for both of the test batteries on each of the tests by sex, race, and total test score. In addition, for the ETB selection was made on type of community, categorizing school districts consisting largely of major cities with high concentrations of population as urban and the remaining districts (small cities, towns, and rural areas) as "other." The analysis of the ETB preceded that of the VGC and, since the effect of types of community was largely negligible in these early analyses and the major concern of the research was with racial bias, the geographic division was not made for the VGC.

The score ranges used to define the samples are given in Table 5. On the ETB, examinees with more than 10% omissions and examinees with perfect scores were eliminated from all tests. Also, since the ETB tests consisted of four-option items, it would be possible to gain a score equivalent to 25% of the items on the basis of purely random guessing. Consequently, examinees below the expected score for completely random guessing were excluded.

On the VGC, the concern was for those examinees who did not make a passing mark, and, thus, the top score range was not examined. Also, the top quartile included largely examinees with nearly perfect scores, whose person-fit statistics would have had very little variance. Medium scorers on the VGC were defined as on the ETB, that is, those examinees falling one half

Table 3: PERSON-FIT STATISTICS FOR RANDOM SAMPLES
EIGHT-TEST BATTERY

<u>Person-Fit Statistics</u>										
	N of cases	N of items	<u>Method A</u>		<u>Method B</u>		<u>Method C</u>		<u>Method D</u>	
			$\bar{x}$	s.d.	$\bar{x}$	s.d.	$\bar{x}$	s.d.	$\bar{x}$	s.d.
Test 1	312	40	35.07	5.70	-0.73	1.46	-0.09	0.81	-1.13	3.24
2	461	20	17.35	3.87	-0.36	1.28	-0.05	0.79	-0.67	2.47
3	480	45	34.76	15.93	-1.17	1.74	-0.11	0.56	-1.62	2.94
4	457	20	18.06	3.33	-0.36	1.10	-0.10	0.79	-0.77	2.29
5	400	48	42.54	12.09	-1.38	2.18	-0.51	0.94	-2.23	4.19
6	481	30	26.75	5.24	-0.65	1.34	-0.26	0.87	-0.97	2.94
7	445	27	21.71	6.57	-0.92	1.25	-0.28	0.62	-1.50	2.37
8	445	40	32.73	6.65	-1.03	1.41	-0.14	0.59	-1.61	3.05

VIRGINIA GRADUATION COMPETENCY TESTS

	N of cases	N of items	<u>Method A</u>		<u>Method B</u>		<u>Method C</u>		<u>Method D</u>	
			$\bar{x}$	s.d.	$\bar{x}$	s.d.	$\bar{x}$	s.d.	$\bar{x}$	s.d.
Reading	365	60	54.33	24.67	-0.07	1.97	0.13	0.43	0.34	2.59
Math	374	99	89.86	44.78	-1.50	3.06	-0.18	0.39	-1.92	4.19

Table 4: INTERCORRELATIONS AMONG PERSON-FIT STATISTICS AND TOTAL SCORES
FOR RANDOM SAMPLES

EIGHT-TEST BATTERY

			Person-Fit Statistics			
	N		Method B	Method C	Method D	Total Test Scores
Test 1: Reading Vocabulary	308	Method A	.70	.47	.48	-.64
		Method B		.93	.89	-.52
		Method C			.88	-.36
		Method D				-.37
Test 2: Reading Comprehension	461	Method A	.65	.47	.48	-.44
		Method B		.96	.94	-.51
		Method C			.91	-.45
		Method D				-.48
Test 3: Reading Essential Skills	480	Method A	.75	.29	.43	-.26
		Method B		.79	.86	-.42
		Method C			.84	-.45
		Method D				-.38
Test 4: Study Skills	457	Method A	.64	.48	.49	-.21
		Method B		.96	.92	-.34
		Method C			.89	-.29
		Method D				-.30
Test 5: Math Computation	400	Method A	.62	.31	.34	.19
		Method B		.91	.88	.26
		Method C			.86	.31
		Method D				.25
Test 6: Math Problem- Solving	481	Method A	.43	.27	.31	-.44
		Method B		.95	.90	-.15
		Method C			.87	.05
		Method D				-.06
Test 7: Math Essential Skills	445	Method A	.76	.43	.53	-.42
		Method B		.88	.89	-.25
		Method C			.84	-.06
		Method D				-.16
Test 8: Occupational Information	445	Method A	.76	.47	.56	-.59
		Method B		.89	.88	-.51
		Method C			.81	-.40
		Method D				-.38

Table 4: INTERCORRELATIONS AMONG PERSON-FIT STATISTICS AND
TOTAL SCORES FOR RANDOM SAMPLES (continued)

VIRGINIA GRADUATION COMPETENCY TESTS

		Person-Fit Statistics			
		Method B	Method C	Method D	Total Test Scores
Reading Test N = 365	Method A	.78	.28	.37	-.03
	Method B		.76	.76	-.02
	Method C			.76	.07
	Method D				-.24
Mathematics Test N = 374	Method A	.80	.25	.40	.09
	Method B		.72	.81	.18
	Method C			.89	.15
	Method D				-.01

Table 5: SCORE RANGES

EIGHT-TEST BATTERY

	N of items	High Score Range	Medium Score Range	Low Score Range
Test 1	40	30-39	20-29	10-19
2	20	16-19	11-15	5-10
3	45	41-44	33-40	11-32
4	20	15-19	10-14	5- 9
5	48	38-47	27-37	13-26
6	30	21-29	13-20	8-12
7	27	22-26	16-21	8-15
8	40	33-39	25-32	10-24

VIRGINIA GRADUATION COMPETENCY TESTS

	N of items	Medium Score Range	Medium-Low Score Range	Low-Low Score Range
Reading	60	45-53	34-44	21-34
Mathematics	99	81-89	70-80	41-69

standard deviation above and below the mean. Low scorers on the ETB fell below the medium range and above the expected random guessing score. This procedure did not suit the VGC test, however, since these easier tests produced such high mean scores that there remained a broad distribution of scores below the medium range. Consequently, for the VGC two levels of low scorers were defined. Medium-low scorers were those who scored below the medium scorers and down to 1.8 standard deviations below the mean. This range roughly represented borderline cases on each test, i.e. those who might be most easily helped to improve their scores to the passing level. The definitions of low-low scorers were based on compromises between frequencies and dispersion.

The four person-fit statistics were computed in each of the 24 cells on each test for the ETB and in the 12 cells for each test on the VGC. To determine if the mean person-fit statistics for cells were significantly different, a series of analyses of variances, using the Statistical Analysis System (SAS, 1976) procedure GLM were performed for each score range with sex, race and community type as independent variables. A summary of the significant effects ($p < .05$) is given in Table 6. Evidence of sex, race, or geographic differences in person-fit statistics among high scorers on the ETB was somewhat scattered. On Test 2 (Reading Comprehension), Test 4 (Study Skills), and Test 8 (Occupational Information), there is consistent evidence of a sex effect, with females showing fewer unusual responses than males. These are also some instances of significant race effects, but the direction of the effect varies, with blacks showing fewer unexpected responses than whites for some tests and some person-fit statistics, and the reverse on other tests.

Among medium scorers on the ETB there are also sex differences on Tests 2 and 4. In addition, there is a reasonably consistent racial effect in Test 1 (Reading Vocabulary), Test 2, Test 4, and Test 8. The main effects for Test 8 are confounded by a significant race by geography interaction, however. On Test 8 white-other examinees and black-urban examinees tend to have the better person-fit statistics. On Tests 1, 2, and 4, the main effects are in a consistent direction with person-fit statistics for whites better than those for blacks, and females better than males. On the VGC, no significant sex or race effects were found for medium scorers.

The most pronounced and consistent evidence of both sex and race effect on the ETB is found for the low scorers across all eight tests. In each case white examinees showed fewer unexpected responses than black examinees and females showed fewer than males.

For medium-low scorers on the VGC, there was a reasonably consistent sex effect for the reading test with females making fewer unusual choices than males. In the single significant race effect for the reading test, blacks had better person-fit statistics than whites.

The main effects for medium-low scorers on the mathematics test were confounded by significant interactions for three of the person-fit measures. In each of these interactions white females had the best person-fit scores

Table 6: ANALYSES OF VARIANCE
EIGHT-TEST BATTERY

High Scorers	Significant* Main Effects			Significant* Interactions			
	S	R	G	SxR	SxG	RxG	SxRxG
Test 1 Method A		X					
B							
C		X	X				
D							
Test 2 Method A	X	X					
B	X						
C	X						
D	X						
Test 3 Method A							
B							
C		X					
D		X					
Test 4 Method A	X			X			
B	X						
C	X						
D							
Test 5 Method A							
B		X					
C		X		X			
D			X				
Test 6 Method A		X			X		
B		X					
C							
D							
Test 7 Method A			X				
B			X				
C							
D							
Test 8 Method A	X						X
B	X						
C	X						
D	X					X	

* $p \leq .05$

Table 6: ANALYSES OF VARIANCE (continued)

EIGHT-TEST BATTERY

Medium Scorers		Significant* Main Effects			Significant* Interactions			
		S	R	G	SxR	SxG	RxG	SxRxG
Test 1 Method	A		X	X				
	B		X					
	C		X					
	D		X					
Test 2 Method	A	X	X					
	B	X	X	X				
	C	X	X	X				
	D	X	X	X				
Test 3 Method	A		X					
	B							
	C							
	D							
Test 4 Method	A	X	X					
	B	X	X					
	C	X	X					
	D	X	X					
Test 5 Method	A	X						
	B	X						
	C	X						
	D							
Test 6 Method	A		X					
	B	X						
	C	X						
	D							
Test 7 Method	A							
	B							
	C							
	D	X						
Test 8 Method	A	X	X					
	B	X	X					
	C							
	D		X					

*p<.05

Table 6: ANALYSES OF VARIANCE (continued)

EIGHT-TEST BATTERY

Low Scorers		Significant* Main Effects			Significant* Interactions			
		S	R	G	SxR	SxG	RxG	SxRxG
Test 1 Method	A	X	X			X		
	B	X	X					
	C	X	X					
	D	X	X					
Test 2 Method	A	X	X					
	B	X	X					
	C	X	X					
	D	X	X					
Test 3 Method	A	X	X		X			
	B	X	X		X			
	C	X			X			
	D	X			X			
Test 4 Method	A	X	X	X				
	B	X	X					
	C	X	X					
	D	X	X					
Test 5 Method	A	X	X					
	B	X	X					
	C	X	X					
	D	X	X					
Test 6 Method	A		X	X				
	B	X	X					
	C	X	X					
	D	X	X					
Test 7 Method	A	X	X	X				
	B	X	X					
	C	X	X					
	D	X	X					
Test 8 Method	A	X	X	X				
	B	X	X	X	X			
	C	X	X	X				
	D	X	X					

*p<.05

Table 6: ANALYSES OF VARIANCE (continued)

VIRGINIA GRADUATION COMPETENCY TESTS

<u>Medium Scorers</u>		<u>Significant[*]</u> <u>Main Effects</u>		<u>Significant[*]</u> <u>Interactions</u>
		<u>S</u>	<u>R</u>	<u>SxR</u>
Reading Method	A			
	B			
	C			
	D			
Math	Method A			
	B			
	C			
	D			
<u>Medium-Low Scorers</u>				
Reading Method	A			
	B	X		
	C	X		
	D	X	X	
Math	Method A			X
	B			X
	C			X
	D			
<u>Low-Low Scorers</u>				
Reading Method	A	X		
	B	X		
	C	X		
	D	X		
Math	Method A	X	X	
	B	X		
	C	X		
	D	X		

* $p \leq .05$

and white males had the poorest of the four groups. Black males and black females were in the middle, with males somewhat better than females.

For the low-low scorers on both the mathematics and reading tests, there were significant sex effects with females consistently better than males. For the single significant race effect, whites were better than blacks.

Item Statistics

The quotient and sum item statistics, formulas (9) and (10), were computed for all samples in this study including the completely random samples. In addition, for the completely random samples, Rasch fit mean squares were computed. Table 7 shows these statistics for the random sample for Test 1-Vocabulary from the ETB. In addition, Table 8 shows the means and standard deviations of the item statistics and their intercorrelations.

These results are typical of most of the tests of this study and are used to illustrate the findings that arose from this area of investigation. Based on the means and standard deviations of the sum and quotient statistics of the random sample, outlier item statistics were identified in each of the 24 cell samples. This goal was accomplished by noting sum and quotient statistics for a sample more than two standard deviations above the mean for the corresponding completely random sample.

For high and medium scoring cells, the result was a hodge-podge. There were a number of items identified as outliers in all of these cells. However, many such items were identified in only one or two cells, and some were identified in all or nearly all cells. No item was consistently identified in cells of any particular category, such as black, white, male or female.

For low scoring cells, many fewer items were identified as outliers, and three of these were identified exclusively for cells containing black examinees. Table 9 shows one of these items and the proportions of low scoring black and white examinees responding to each choice. Though about the same proportion of each group answered correctly, there were substantial differences in popularity of the wrong choices. Moreover, there is no indication that black (or white) examinees were guessing largely at random in the absence of knowledge, as would have been indicated by nearly equal proportions selecting each distractor. Instead, it appears that for some reason, which may remain forever unexplained, white examinees were especially prone to choose "consider" as a synonym for "assume," while black examinees were approximately equally likely to choose "consider" and "avoid."

Identification of items reflecting racial or sex differences was accomplished much more often by use of the sum statistic than by use of the quotient statistic. There were a number of items identified in all or nearly all the cells of a category (e.g., low-scoring female) by the sum statistic

Table 7: SUM AND QUOTIENT STATISTICS AND RASCH FIT MEAN SQUARES
FOR TEST 1 RANDOM SAMPLE

EIGHT-TEST BATTERY N = 312

Item No.	Sum	Quotient	Rasch Fit MS
1	-0.052	1.404	0.81
2	-0.052	1.093	0.55
3	-0.034	3.612	0.62
4	-0.025	1.743	0.99
5	-0.042	1.805	0.84
6	-0.040	1.500	0.71
7	-0.010	1.811	0.99
8	-0.047	1.681	0.83
9	-0.043	1.656	0.69
10	-0.033	1.719	0.88
11	-0.029	1.495	0.89
12	-0.034	3.066	0.74
13	0.008	1.802	1.16
14	-0.017	1.611	0.79
15	-0.025	1.387	0.83
16	-0.029	1.633	0.92
17	-0.016	1.776	0.74
18	-0.004	3.065	1.16
19	-0.068	1.030	0.47
20	0.017	2.098	1.05
21	0.004	1.492	0.96
22	-0.026	1.962	0.75
23	-0.039	1.388	0.68
24	-0.022	1.462	0.82
25	0.016	2.009	1.02
26	0.062	6.659	1.26
27	-0.031	1.418	0.88
28	0.035	2.161	1.27
29	-0.005	1.552	0.88
30	0.007	1.493	1.08
31	-0.014	1.427	1.00
32	0.028	2.306	1.09
33	0.024	2.587	1.47
34	0.045	1.992	1.28
35	0.006	3.516	1.19
36	0.060	4.132	1.48
37	0.052	2.526	1.45
38	0.018	2.036	1.21
39	0.088	4.075	2.61
40	0.014	1.969	0.96

Table 8: MEANS, STANDARD DEVIATIONS AND INTERCORRELATIONS FOR SUM AND QUOTIENT STATISTICS AND RASCH FIT MEAN SQUARES FOR TEST 1 EIGHT-TEST BATTERY

	N of Items	Mean	s.d.	Correlations		
				Sum	Quotient	Rasch
Sum	40	-0.01	0.04	--	.63	.86
Quotient	40	2.13	1.06	--	--	.53
Rasch Fit Mean Square	40	1.00	0.36	--	--	--

Table 9: RESPONSES OF LOW-SCORING BLACK AND WHITE EXAMINEES TO AN EXAMPLE OUTLIER ITEM ON TEST 1 EIGHT-TEST BATTERY

Item No. 26	<u>undertake</u> the job			
Options:	<u>consider</u>	<u>avoid</u>	<u>betray</u>	<u>assume*</u>
White Proportion N = 455	.40	.24	.19	.17
Black Proportion N = 445	.33	.30	.20	.18

*correct response.

and not simultaneously by the quotient statistic. The reverse was not observed for any test; that is, the quotient statistic did not identify items related to specific groups independently of the sum statistic. Very few group specific items could have been identified by the quotient statistic compared with the number identified by the sum statistic. This statement does not imply that the quotient statistics were not two standard deviations above their means as often as the sum statistics. Instead, the result was that outlying quotient statistics were rarely noted consistently for the same item over the samples representing a category, such as low scoring black examinees.

In summary, the following may be said about the items identified as especially contributing to lack of person-fit for specific categories of examinees:

- 1) There were relatively very few such items, and these were associated almost exclusively with low-scoring groups.
- 2) There was no reasonable explanation as to why examinees in any category answered certain items in a way that tended to generate poor person-fits.
- 3) The response distributions for items which were associated with poor person-fits could not be interpreted as contributing to score bias.

Tables 7 and 8 were presented as characteristic of most of the tests of the study. This qualification was necessary because for two tests, the pattern of intercorrelations among the item statistics was different from that shown in Table 8. These tests were Test 5-Math Computation and the mathematics test of the VGC. For these two tests, the Rasch fit mean squares correlated somewhat negatively with the sum statistics and positively with the quotient statistics. In contrast, for all other tests, the Rasch statistic correlated positively with both the sum and quotient statistics but somewhat more positively with the sum statistic. In spite of this reversal, earlier comments about greater effectiveness of the sum statistic for identifying items contributing to poor person-fit statistics for specific groups hold for the two mathematics tests as well.

Sure Knowledge and Dispersion Indices

Two additional item statistics were combined for all items on a given test, to yield summary measures. Using the Horst (1966) method described above, the number of items known with assurance was computed for black and white examinees and male and female examinees by score level by test. The percentage of knowledge for each group on each test was determined by a ratio of the estimated mean number of items known with assurance to the mean number of items answered correctly.

The second summary measure was based on the index of dispersion (11). These indices were computed for each item on a test, summed across items, and then divided by the number of items on a given test. This division had the effect of standardizing the indices for comparison across tests.

Although the analyses of the person-fit statistics suggested that black examinees were more likely to select unusual responses than white, and males more likely than females, the hypothesis that this tendency was the result of poor use of partial information was not borne out by an examination of the relative portions of items known with assurance for these groups in terms of their total scores. Two examples of the analysis of knowledge and dispersion, a reading test from each of the test batteries, are shown in Table 10. Overall, within ability groups white examinees consistently knew with assurance a larger percentage of the items which they answered correctly than did black examinees as determined by the Horst procedure. Females showed the same superiority in comparison to males. These results suggest that rather than using partial information poorly, black examinees and male examinees were using their partial information with considerable skill to attain the same scores as white examinees and female examinees in the same ability groups. The effect is most pronounced among the lower scoring examinees and is stronger on the ETB than on the Virginia tests.

The mean dispersion indices were intended to measure the relative degree of uniformity of the proportions of choices among the wrong responses. The lower indices indicated a greater tendency towards random guessing. As is evident from Table 10 and from the analyses of the remaining tests, there was little difference within ability groups in the amount of random guessing. Where differences appear, they actually represent only one or two percentage points shift in responses from one option to another and, overall, indicate negligible differences.

Discussion

Characteristics of Person-Fit Measures

The four person-fit measures of the study were substantially and significantly intercorrelated in every sample for every test. Moreover, when significant mean differences in person-fit were observed, usually all four measures were significant. This outcome suggests that, in a sense, all four measures relate to the same phenomenon. However, the Rasch person-fit statistic, Method A, seemed weaker in this regard for the more difficult ETB. This conclusion is based on the magnitude of the F ratios when three or four of the methods yielded significant race or sex effects or (in one analysis) a significant sex by race interaction. The 64 analyses of variance conforming to these outcomes permitted 24 comparisons of sets of four F ratios. In only three of these did Method A result in the highest F ratio. In contrast, for the easier VGC, there was no obvious difference in performance among the four methods. This outcome is plausible, because of a preponderance of easy items on the Virginia tests. As noted above, the Rasch

Table 10: ITEMS KNOWN WITH ASSURANCE AND MEAN DISPERSION INDICES
EIGHT-TEST BATTERY Test 3 - Reading Essential Skills (N of items = 45)

	N of Examinees	Mean No. of Known Items	Mean % of Knowledge	Mean Dispersion Index	N of Excluded Items
High White	462	40.86	96.1	.013	0
High Black	315	40.02	95.2	.016	0
High Male	344	40.59	95.7	.014	0
High Female	433	40.42	95.6	.014	0
Med. White	404	33.54	89.0	.030	2
Med. Black	435	31.61	87.2	.029	2
Med. Male	421	32.63	88.4	.029	2
Med. Female	418	32.50	87.9	.030	2
Low White	423	17.72	68.9	.035	5
Low Black	428	14.24	61.0	.038	6
Low Male	421	14.37	61.8	.037	5
Low Female	430	17.58	68.3	.038	6
Random Sample	480	31.43	87.4	.025	0

VIRGINIA GRADUATION COMPETENCY TESTS

Virginia Reading Test (N of items = 60)

	N of Examinees	Mean No. of Known Items	Mean % of Knowledge	Mean Dispersion Index	N of Excluded Items
Med. White	283	43.04	87.6	.049	2
Med. Black	271	41.68	86.3	.047	2
Med. Male	226	42.14	86.7	.045	2
Med. Female	328	42.59	87.3	.045	3
Med. Low White	307	28.63	72.5	.065	5
Mws. Low Black	230	28.33	71.8	.072	6
Med. Low Male	267	28.60	72.5	.063	7
Med. Low Female	270	28.46	71.9	.073	6
Low Low White	163	13.88	47.1	.093	15
Low Low Black	204	13.47	47.2	.094	16
Low Low Male	196	13.41	45.9	.093	16
Low Low Female	171	14.28	48.2	.094	15
Random Sample	365	42.67	87.4	.042	2

person-fit statistic (Method A) is based only on differences in right answer versus wrong answer probabilities, while the other three person-fit statistics are determined from the probabilities associated with each choice on an item. On very easy items, the difference in the distribution of small probabilities of being wrong among three options as opposed to one option would be slight, whereas on more difficult items the lumping together of all wrong probabilities (as in the Rasch statistic) could result in a substantial difference from the other three methods.

Person-fit statistics from Methods B and D were highly intercorrelated. This outcome was to be expected, since each of these methods effectively weights each test item equally in summing to determine the person-fit statistic. Though they standardize the fit deviation for each item differently, apparently the outcomes are approximately equivalent.

Method C weights the test items according to raw deviation from expectation. Those items on which the examinee's responses deviate most from expectation play a proportionately much greater role in establishing person-fit than they would in Methods B and D. Despite this difference Method C person-fit statistics correlated very substantially with those of Methods B and D. Also, the Method C statistics yielded about the same pattern of significant differences among groups.

All of the person-fit statistics diverged somewhat from theoretical expectations concerning their distributions. Means were generally lower than suggested by theory, but not always and not greatly so. However, standard deviations were substantially smaller than under theory. Both of these outcomes could be due to inaccuracy of the probability estimates used. This inaccuracy, in turn, may have been due to inappropriateness of the assumption of linear relationships between observed score and probabilities of selecting choices. Trials using other functions might resolve this question. However, for practical purposes, a successful resolution of the question might not be particularly helpful, since the person-fit statistics as calculated appeared to reflect group differences consistently over a wide range of sampling.

Person-Fit Outcomes

There can hardly be a question that within the populations of this study there are real person-fit differences between some subgroups of male and female and black and white examinees. These differences were particularly apparent among low scoring groups, but even here were more pronounced for some tests than others.

For example, for Test 1 (Vocabulary) of the ETB, white-black differences in person-fit for low scorers averaged about twice as large as corresponding male-female differences. For Test 2 (Reading Comprehension), the black-white and male-female person-fit differences for low scorers were about equal in magnitude. The significant interaction for Test 3 (Reading Essential Skills) reflected low-scoring black male person-fit statistics

much higher than for corresponding black females or white examinees of either sex. These and many other observed variations in pattern would seem to be test specific.

This judgement is further borne out by an interesting reversal of magnitude for person-fit statistics as related to raw test scores for mathematics computation. This reversal occurred both for the mathematics computation test within the ETB and for the VGC mathematics test. For all tests analyzed other than these two, high (poor) person-fit statistics were associated with low scorers. This outcome is plausible because high scorers are most likely to select right answers and have relatively few chances to do otherwise. Nevertheless, for both largely computational mathematics tests, the worst person-fits were attained (on the average) by high scorers and the best by low scorers. This outcome would suggest that high scorers, when they did miss an item, did so for reasons other than lack of knowledge and in so doing often selected highly improbable wrong choices. Conversely, low scorers tended to select popular wrong choices. This analysis is plausible because many wrong choices on mathematics computation tests are the results examinees obtain through common-place errors. Low scorers would be prone to selecting these wrong answers, which would also be those chosen most prevalently. In contrast, better examinees would seldom make such errors and hence would be more likely to select unpopular wrong answers.

For middle-score samples of the ETB significant race and sex differences in person-fit were found in somewhat the same patterns as for low-scoring groups. However, the differences in magnitude of observed mean person-fits were quite small compared with those between low-scoring groups. Significant differences among high scoring groups on the ETB were scattered and inconsistent in direction and were usually associated with much higher probabilities than corresponding tests for middle and low scoring groups. Thus it is reasonable to conclude that these significant F ratios represent inconsequential differences or type I errors.

Person-fit results were somewhat different for the VGC tests. There were many fewer instances of group differences. Significant race differences were not observed systematically across the four person-fit statistics at any score level. Sex differences were somewhat consistently significant for the Virginia reading test, and these were of higher magnitude than for most significant sex differences for the ETB. There was a significant sex by race interaction for low-low scorers on the Virginia math test, but this outcome can hardly have practical implications given the reversal of the relationship between total score and goodness of person-fit.

Because these outcomes were partly at variance with those of the ETB (especially with respect to race differences), several additional analyses were made which are not covered in the preceding section of this report. One analysis broke down the Virginia reading test into subtests, one of which closely resembled a test from the ETB. It was believed possible that the somewhat heterogeneous nature of the Virginia reading test might have contributed to the sparse outcomes associated with it. None of the analyses of

the Virginia reading subtests yielded results substantially different from those for the entire test. If anything there were fewer significant comparisons, which may have been due to the shortness of the subtests.

Considering the foregoing discussion, it seems clear that the VGC tests do differ from those of the ETB with respect to person-fit. This difference may be a function of the different and easier character of the Virginia tests.

Apparent Causes of Mean Person-Fit Differences - Item Statistics

One possible cause of person-fit differences would be the presence of moderate numbers of biased items, that is, items relatively easy for one group and difficult for another. It was for this reason that item statistics were generated for all samples of the study. From the random sample results for each test, it is clear that the sum statistic of this study, formula (10), is substantially related to the Rasch fit mean square.

The sum statistic identified a relatively small number of items contributing somewhat more so than others to poor person-fit statistics for examinees in certain categories. Almost all of these items were identified with respect to low-scoring groups. However, inspection of the items themselves and the response distributions for the groups in question yielded no suggestion of actual bias. Instead, the usual circumstance was that a substantial proportion of the group with worse person-fit preferred an unpopular wrong choice while the proportion selecting the correct choice was about the same as for contrasting groups. While this phenomenon accounted for some of the difference in mean person-fit statistics between groups, it was observed far too infrequently to be a major contributing factor.

The quotient statistic, while slightly to moderately correlated with the sum statistic, yielded much less consistent results. It identified no items contributing to poor person-fit for a specific group of examinees independently from the sum statistic. This performance may be due to inherent instability of the statistic. This instability may arise from the exaggerated effect of an occasional very small probability of selecting a response coupled with sampling outcomes yielding slight excesses of responses on the corresponding choices.

The sum statistic is recommended as a possibly more powerful and sensitive item statistic than the Rasch fit mean square. This recommendation is not based on direct comparison of the performance of the two statistics but is based on the greater amount of information going into the sum statistic and its performance in identifying items contributing to mean person-fit differences between groups.

Regardless, the small number of items so identified could hardly have accounted for a major proportion of the observed differences in person-fit between some groups. Apparently this effect was cumulative across a large proportion of the items on tests for which significant differences were observed.

Apparent Causes of Mean Person-Fit Differences - Dispersion Indices

Differences in dispersion indices could easily explain person-fit differences between groups. For example, random guessing in the absence of knowledge would yield a low dispersion index, while extensive elimination of implausible wrong choices would result in a higher index.

These indices were averaged across items for each test of the study at each score level separately for males and females and for black and white examinees. The results yielded no explanation for observed mean person-fit differences. For low scorers on the ETB, four tests yielded greater mean dispersion indices for black examinees and four for white examinees. (Recall that low-scoring, black-white person-fit differences for all of these tests were statistically significant.) Moreover, the observed low-scoring, black-white differences in mean dispersion indices were very small, typically about .002.

Male-female comparisons of mean dispersion indices for low scorers on the ETB indicated that females' responses were consistently more dispersed than males. Female responses yielded higher mean dispersion indices on all eight tests, but the differences were small, only .01 on the average.

Male-female and black-white comparisons at other score levels yielded similar results for the ETB. For example, mean dispersion indices for high-scoring black examinees were higher than for corresponding white examinees on seven of the eight tests, but only to a very slight degree. The reverse situation prevailed for medium scorers, with whites attaining slightly higher dispersion indices on six of the eight tests.

Dispersion indices from the VGC tests also provided no explanation for observed mean person-fit differences. Between-group differences at each score level were very small.

Apparent Causes of Mean Person-Fit Differences - Items Known With Assurance

The difference between the proportion answering an item correctly and the proportion selecting the most popular distractor may well be questioned as an adequate measure of the proportion knowing the answer with assurance. While the model may yield inaccurate estimates for some specific items, summing the estimated sure knowledge proportions across all items should give a reasonable estimate of the average total proportion known with assurance by the group. Sometimes the model may overestimate and sometimes underestimate, but over all items there should be some balancing of these effects. A likely instance of underestimation would be an item for which the answer was not the most popular response. Then use of the model yields an estimated proportion with sure knowledge of zero. Thus it might be more realistic to view the sum of the sure knowledge proportions as representing only the items of the test for which the answers were the most popular responses.

In this case, group comparisons of the proportion of sure knowledge must take into account the number of items disqualified on this basis.

For the analyses of this study, especially for low-scoring groups, there were a number of items which were disqualified or for which it would be necessary to conclude that no examinees had sure knowledge. However, for all group comparisons, there were about an equal number of such items, and largely the same items were disqualified for both groups. Therefore, in spite of potential weaknesses, the basis adopted for estimating the proportion of sure knowledge was judged adequate.

The sure knowledge analyses provide the only outcomes of the study consistent with the person-fit outcomes. Specifically, the significant person-fit differences for sex and race coincide exactly with rather substantial differences in the estimated proportion of sure knowledge. It would appear that the worse person-fit statistics for black and male examinees within a score group are strongly associated with lower levels of sure knowledge. Necessarily then, these black and male examinees gained more from partial information or guessing after elimination some wrong choices than did their white or female counterparts.

This outcome is entirely consistent with the observed person-fit differences between groups. Examinees with less sure knowledge would be guessing more so than others on items of every difficulty level. Some of these items would be easy, with low probabilities associated with all wrong choices even for relatively low scoring examinees. Those guessing on such items would inflate their person-fit statistics greatly whenever a wrong answer was chosen. In contrast, those with greater sure knowledge would have this knowledge concentrated most heavily among easier items and more rarely select a distractor of these items.

Bias Due to Omission of Responses

Though study of omissions was not a specific goal of this study, this phenomenon imposed itself upon the process. It became necessary to delete small proportions but numerically large numbers of examinees from analyses of the ETB due to excessive numbers of omissions. Such examinees if selected for any random sample would, *per se*, have been found to have very poor person-fit statistics for Methods B, C, and D due to low probabilities for omissions at any score level. (An omission is categorized as one of the possible responses to an item in these methods.)

For the Virginia data, only examinees with no omissions were used in constituting the (otherwise) random samples. For these tests, there were so few omissions that even one or two omissions by an examinee might have resulted in an extreme person-fit statistic.

Inspection of Table 2 reveals that in proportion to their numbers many more black than white examinees were dropped from the population for the ETB.

Since there was no penalty or admonition of any kind against guessing on these tests, failure to attempt all items is prima facie evidence of behavior which, independent of knowledge, depresses scores. Since this behavior is more prevalent among black examinees than white, the tests are, in a sense, biased against black examinees. For the Virginia tests also, substantially more black than white examinees were excluded for omissions (14% vs. 7%), but the effect on scores was minimal due to the small number of omissions for a typical excluded examinee.

Conclusions

At first glance the inflated person-fit statistics for some subgroups of black and male examinees seems strongly suggestive of bias against these groups. However, poor person-fit statistics are neutral in themselves. They merely attest to a tendency to select improbable responses, improbable largely in the sense that they are not the responses more popular for the population of examinees. Such a tendency may or may not cause the examinee to earn an unfairly low score. If the inflated person-fit statistic arises from failure to eliminate obvious distractors, unfairness would probably result. In contrast, strong preferences for unpopular distractors over very popular ones in the absence of knowledge would not necessarily result in an unfairly low score.

For the tests of this investigation, there was no confirmable evidence of bias against blacks or males associated with person-fit differences, though for certain tests and score groups male and black person-fit statistics were significantly worse than for their female and white counterparts. This phenomenon was found for all of the low-scoring groups on the ETB, but only for very low-scoring males on the VGC. It would appear that the poor person-fit statistics for some groups of black and male examinees reflected, instead of bias against them, their superior ability to gain score points by effective guessing in the absence of as much sure knowledge as possessed by their white and female counterparts whose total scores were no higher.

At first glance, one might be tempted to conclude that, therefore, there was bias against some subgroups of white or female examinees; however, such a claim cannot be justified given the nature and purpose of the tests analyzed. The ETB is a typical standardized achievement battery. The use of scores from this battery is fully consistent with reward of partial information. The very norms of such tests are based on assuming that, across the entire score range of the test, different mixes of partial and complete information have the same value if they yield the same score. In contrast, a minimum competency test is designed to simulate real-life tasks and has a fixed cutoff in the center of a narrow score range of interest. The presence of unusual guessing success for some groups with scores in this range could have indicated, in a sense, bias in favor of these groups. However, for the VGC, the only consistent evidence of this sort was for very low-scoring groups of males, who could not have gained enough from guessing to pass the tests.

While the methodology specifically developed for this study revealed no evidence of bias in the tests analyzed, it has the potential to do so as evidenced by the consistent and revealing performance of the measures taken. This outcome is certainly the most important of the study. Also important is the finding that no substantial evidence of racial bias or even poor person-fit across racial groups could be found for a large-scale, statewide competency testing program. This outcome and the apparent lack of bias due to person-fit differences in the ETB do not necessarily imply that the overall outcome of such testing programs is free from bias or unfairness. Indeed, the tendency for black examinees to omit more items than white examinees is clear evidence of bias. However, this aspect of bias affected only a small proportion of black examinees and was noticeably minimal in its consequences for the VGC. Thus, for the VGC and ETB, future investigations of score differences between black and white students might more profitably concentrate on factors external to testing, such as inequality of educational treatment.

References

- Barr, A. J., Goodnight, J. H., Sall, J. P., & Helwig, J. T. A User's Guide to SAS 76. Raleigh, North Carolina: SAS Institute Inc., 1976.
- Cleary, T. A., & Hilton, T. L. An investigation of item bias. Educational and Psychological Measurement, 1968, 28, 61-75.
- Cole, N. S. Bias in selection. Journal of Educational Measurement, 1973, 10, 237-255.
- Echternacht, G. A quick method for determining test bias. Educational and Psychological Measurement, 1974, 34, 271-280.
- Frary, R. B., Tideman, T. N., & Watts, T. M. Indices of cheating on multiple-choice tests. Journal of Educational Statistics, 1977, 2, 235-256.
- Good, I. J. The surprise index for the multivariate normal distribution. The Annals of Mathematical Statistics, 1956, 27(4), 1130-1135.
- Hambleton, R. K., et al. Developments in latent trait theory: A review of models, technical issues, and applications. Paper presented at the joint meeting of the National Council on Measurement in Education and the American Educational Research Association, New York, April, 1977.
- Horst, P. Psychological Measurement and Prediction. Belmont, California: Wadsworth Publishing Company, Inc., 1966.
- Lane, C. E. A comparison of the effects of conventional testing and two-stage testing procedures on item bias as defined by three statistical techniques. Unpublished doctoral dissertation, Virginia Polytechnic Institute and State University, 1978.
- Linn, R. L., & Werts, C. E. Considerations of test bias. Journal of Educational Measurement, 1971, 8, 1-4.
- Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., & Bent, D. H. Statistical Package for the Social Sciences (2nd ed.). New York: McGraw-Hill Book Company, 1975.
- Weaver, W. Probability, rarity, interest and surprise, Scientific Monthly, 1948, 67, 390-392.
- Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-115.
- Wright, B. D., Mead, R. M., & Draba, R. Detecting and correcting test item bias with a logistic response model (Research Memorandum No.22), Chicago: Statistical Laboratory, Dept. of Education, University of Chicago, 1976.